

# ANALYSIS OF IMAGES SOCIAL NETWORKS AND TEXTS

## Bigram Anchor Words Topic Model

Ashuha Arseniy, Loukachevitch Natalia

Moscow Institute of Physics and Technology  
Research Computing Center of Lomonosov Moscow State University

`ars.ashuha@gmail.com louk_nat@mail.ru`

April 29, 2016

## Motivation

- ▶ Nowadays we have a lot of data, but usually **it is unlabeled**
- ▶ We want to extract structure from document collection **unsupervised**

## How can we get this goal?

- ▶ Topic modeling is a powerful tool for document collection analysis
- ▶ Informally, topic is a semantically related set of words

sample: *geom rna fast dna sequence alignment nucleotides*

## More formal

- ▶ Topic is a discrete distribution over words  $p(w|t) = p(\text{word}|\text{topic})$
- ▶ Document is a discrete distribution over topics  $p(t|d) = p(\text{topic}|\text{doc})$
- ▶ We want to find  $p(w|t)$ ,  $p(t|d)$  given  $p(\text{word}|\text{doc})$
- ▶ Usually we solve this problem as a matrix decomposition

## Topics

gene 0.04  
dna 0.02  
genetic 0.01  
...

life 0.02  
evolve 0.01  
organism 0.01  
...

brain 0.04  
neuron 0.02  
nerve 0.01  
...

data 0.02  
number 0.02  
computer 0.01  
...

## Documents

## Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here, "two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**." One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden. "We arrived at the 800 number. But coming up with a consensus answer may be more than just a **few** numbers since, particularly for more and more **genomes** are completely sequenced and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



**Stripping down.** Computer analysis yields an estimate of the minimum modern and ancient genomes.

## Topic proportions and assignments

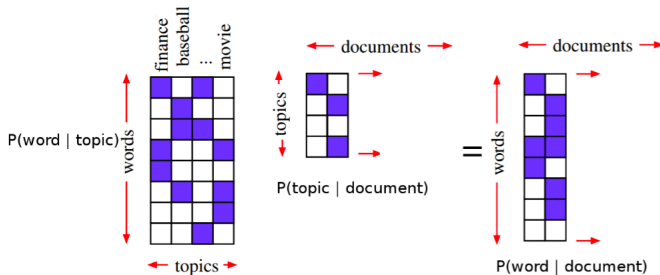


## Probabilistic model

$$p(\text{word}|\text{doc}) = \sum_{\text{topic}} p(\text{word}|\text{topic})p(\text{topic}|\text{doc})$$

- ▶ The order of words in document is not matter (bag of words)
- ▶ Topic is not depends on doc ( $p(\text{word}|\text{doc}, \text{topic}) = p(\text{word}|\text{topic})$ )

Represent it as **matrix decomposition** and solve this problem by MLE

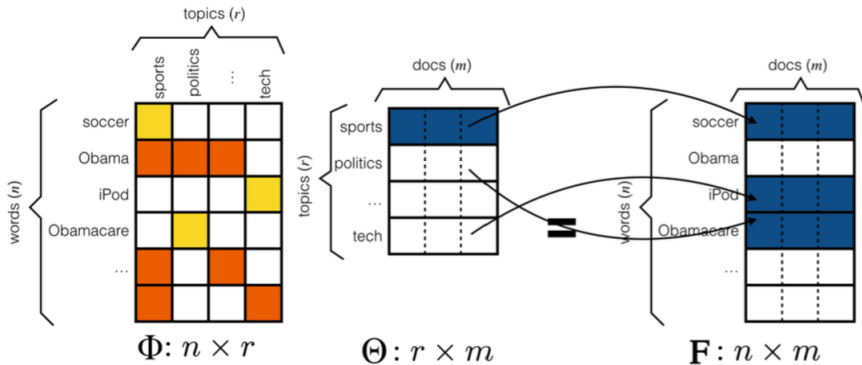


## Regularization

- ▶ LDA – topics and documents generated from Dirichlet distribution
- ▶ BigARTM – generalize LDA, many regularizes

- + Good matrix approximation
- + A lot of implementations
- + There exist modification to take into account bigrams
- Solution is really depends on initial approximation
- Poor model of documents
- Difficult to parallelize
- Computational difficult
- Control coefficient of regularizations is really hard task

Let's assume for each topic  $T$  there exist word  $w$  that  $p(w|t) \neq 0$  if  $t = T$



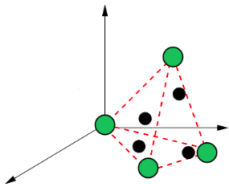
Therefore  $F$  is just a linear composition constructed  $\Theta$  rows, anchor rows.

1. How can we find rows in  $F$  which correspond to anchor words?
2. How can we reconstruct topic model  $(\Phi, \Theta)$  given anchor words?

- ▶ Matrix  $F$  too noisy, let's use  $FF^t$
- ▶ Size of  $FF^t$  is  $Words \times Words$  therefor reduce dimension

$$FF_{words \times words}^t = H_{words \times k}$$

- ▶ Find almost convex hull in rows of  $H$  matrix  $\{H_{anchor_1}, \dots, H_{anchor_n}\}$



- ▶ Solve **independent** convex optimization problems: find  $c_i$  for each  $t$

$$H_t \approx \sum_{i=1, \dots, T} c_{ti} H_{anchor_i}, \quad c_{it} \geq 0, \quad \sum_i c_{it} = 1, \quad c_i = p(topic|word)$$

- ▶ Use Bayes rule to reconstruct  $\Phi = (p(word|topic))_{W \times T}$

- + No initial approximation
- + Very well parallelize out of box
- Need to tune parameters
- Can't take into account bigrams
- Worst matrix decomposition

Our goal was propose modification witch **can take into account bigrams**.

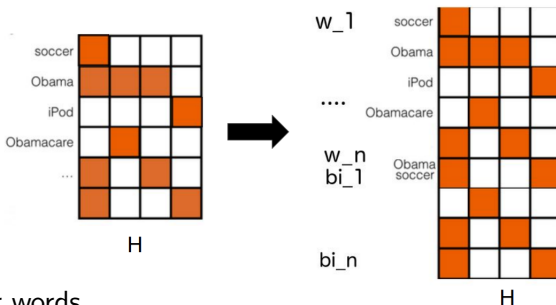
Why it is important?

- ▶ adding new information about word order in model
- ▶ better and lighter interpretability
- ▶ simple solution – does not work



There is one simple way:

1. precomputed bigrams
2. we assume that vector for bigram  $w_i w_j = H_{w_i} + H_{w_j}$
3. add vectors corresponds bigrams to set of points H (finding anchors)



4. Find anchor words
5. Recover topic mode
6. Make some PLSA steps

**Bigrams** can be anchor words

**Interpretations** good latent space in matrix  $H$

## Old anchors

- ▶ loss
- ▶ cluster
- ▶ mixtur
- ▶ synaps
- ▶ theorem
- ▶ speech
- ▶ entropi
- ▶ filter
- ▶ competit
- ▶ gain
- ▶ markov
- ▶ identif
- ▶ algorithm




## Our anchors

- ▶ mixtur
- ▶ boltzmann\_machin
- ▶ likelihood
- ▶ markov\_chain
- ▶ action
- ▶ vector\_quantiz
- ▶ network
- ▶ robot\_arm
- ▶ loss
- ▶ tangent\_distanc
- ▶ classifi
- ▶ reinforc\_learn
- ▶ speech

## Metrics:

- ▶ **Perplexity** is a mean  $\exp(-\text{mean likelihood})$
- ▶ **Coherence** is a mean Pointwise Mutual Information
- ▶ **Unique of kernels** is a mean Jaccard distance between most probable words in topic

Collection	Banks Articles			20 Newsgroups			NIPS		
Metric	$P_{test}$	PMI	$U$	$P_{test}$	PMI	$U$	$P_{test}$	PMI	$U$
PL	2116	0.60	0.40	2155	0.31	0.40	1635	0.21	0.32
AW	2330	0.63	0.53	2268	0.38	0.41	1505	0.41	0.38
BiAW	2248	0.79	0.60	2183	0.68	0.54	1500	0.50	0.41
AW+PL	2052	0.78	0.58	2053	0.54	0.55	1434	0.52	0.46
BiAW+PL	<b>1848</b>	<b>0.87</b>	<b>0.63</b>	<b>2027</b>	<b>0.78</b>	<b>0.64</b>	<b>1413</b>	<b>0.58</b>	<b>0.49</b>

-  Arora, S., Ge, R., Moitra, A.: Learning topic models - going beyond svd. In: Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on, IEEE (2012) 1–10
-  Arora, S., Ge, R., Halpern, Y., Mimno, D., Moitra, A., Sontag, D., Wu, Y., Zhu, M.: A practical algorithm for topic modeling with provable guarantees. arXiv preprint arXiv:1212.4777 (2012)
-  B, Dobrov, N, Loukachevitch: Forming the base of terminological phrases in the texts of the subject area (2003) 201–210